A Hybrid Approach for Customer Segmentation and Loyalty Prediction in E-Commerce

Elamurugan Balasundaram [1]
P. Aranganathan [2]
Krishna Sudhir Annavajjala [3]
R. Sivakumar [4]
Mathiazhagan Arumugam [5]
A. Vinoth [6]

Abstract

Purpose : This study aimed to enhance e-commerce customer segmentation and loyalty prediction by integrating machine learning (ML) with traditional statistical methods.

Design/Methodology/Approach : The research adopted a hybrid approach, utilizing k-means clustering for customer segmentation based on recency, frequency, and monetary values, followed by an XGBoost classifier application for loyalty prediction. The methodology involved analyzing actual e-commerce data and comparing results with established industry benchmarks.

Findings : The hybrid model demonstrated superior performance over conventional methods, evidenced by improved precision, recall, accuracy, and *F*1 scores in loyalty prediction, alongside higher silhouette scores and lower Davies–Bouldin indices for customer segmentation.

Practical Implications : The approach provided a more generalized, interpretable, and high-quality framework for e-commerce businesses to understand customer behavior and enhance retention strategies.

Originality/Value : The research contributed to the field by presenting a novel method that successfully combines ML and statistical analysis, offering a more effective solution for customer segmentation and loyalty prediction in e-commerce settings.

Keywords : **customer segmentation, loyalty prediction, e-commerce, *k*-means clustering, XGBoost**

JEL Classifications Codes : C38, C45, C55, L81, M31

Paper Submission Date : September 20, 2023 ; Paper sent back for Revision : May 24, 2024 ; Paper Acceptance Date : July 15, 2024 ; Paper Published Online : October 15, 2024

[1] *Associate Professor,* Department of Management Studies, Sri Manakula Vinayagar Engineering College, Madagadipet, Mannadipet Commune - 605 107, Puducherry. (Email : harshadelamurugan@gmail.com)
ORCID iD : https://orcid.org/0000-0002-9747-1727
[2] *Associate Professor,* Gnanam School of Business, Mary's Nagar, Trichy-Thanjavur Express Highway, Sengipatty, Thanjavur - 613 402, Tamil Nadu. (Email : aranganathanp@gmail.com)
ORCID iD : https://orcid.org/0000-0002-4132-5704
[3] *Professor & HOD,* Department of MBA, Koneru Lakshmaiah Education Foundation, Green Fields, Vaddeswaram, Guntur District, Andhra Pradesh. (Email : krishnasudheer.a@klh.edu.in)
[4] *Assistant Professor,* Department of Management Studies, Sri Manakula Vinayagar Engineering College, Madagadipet, Mannadipet Commune - 605 107, Puducherry. (Email : shiva12may@gmail.com)
ORCID iD : https://orcid.org/0000-0002-1237-3977
[5] *Assistant Professor,* Department of Management Studies, Sri Manakula Vinayagar Engineering College, Madagadipet, Mannadipet Commune - 605 107, Puducherry. (Email : mathiazhagan.a@gmail.com)
[6] *Assistant Professor,* Department of Management Studies, Sri Manakula Vinayagar Engineering College, Madagadipet, Mannadipet Commune - 605 107, Puducherry. (Email : avinubemba@gmail.com)

DOI : https://doi.org/10.17010/pijom/2024/v17i10/173996

The e-commerce industry in India is experiencing rapid growth, with market projections indicating a worth of $74.8 billion by 2031 (trade.gov, 2022). This sector is expected to grow the fastest within the Asia-Pacific region, attracting significant investments from venture capitalists and corporate giants and anticipating 100% foreign direct investment (FDI) (Joshi & Achuthan, 2016). Consequently, e-commerce business models have garnered substantial interest among researchers and academicians (Dhote & Zahoor, 2017). To remain competitive in the face of escalating market competition and customer expectations, e-commerce companies must execute effective tactics to keep existing customers while attracting new ones (Nisar & Prabhakar, 2017). Two critical tasks that support these objectives are customer segmentation and loyalty prediction (Adelaar et al., 2004; Shobana et al., 2023). However, segmenting customers and predicting loyalty poses significant challenges due to the need to manage massive, high-dimensional, noisy, and dynamic data (Agrawal et al., 2023; Ahmed & Kumari, 2022). Current techniques primarily rely on traditional statistical methods or machine learning (ML) models (Andersson & Börjeson, 2023). Recency, frequency, monetary (RFM), *k*-means clustering, and logistic regression analysis are classic methods with performance limitations, such as the reliance on and need for feature engineering and over-sensitivity to outliers (Agrawal et al., 2023; Nisar & Prabhakar, 2017; Rahayu et al., 2022). ML models are not free from shortcomings; for instance, overfitting, model underfitting, and large volumes of labeled data are needed (Costa & Pedreira, 2023; Huyut & Üstündağ, 2022; Kumar et al., 2016; Kushwah et al., 2022).

We have suggested the use of hybrid methods like these but they do not present adequate frameworks addressing practically these deficiencies. For instance, Lee and Jiang (2021) merged supervised classification and unsupervised clustering, but the model needs further testing on other datasets. Concurrently, there is equally a gap in the perspective of developing methods that integrate prediction and segmentation in one system (Ullah et al., 2023; Wu et al., 2022). As a result, it is possible to conclude that addressing this gap necessitates the development of a focused parameter-defining and multi-level customer segmentation method that employs both classic and sophisticated ML techniques.

This research aims to fill the gap by creating a hybrid method that combines statistical and ML techniques for e-commerce customer segmentation and loyalty prediction. Our proposed strategy has two main parts: customer segmentation using RFM values with *k*-means clustering and customer loyalty prediction using demographic and behavioral features with the XGBoost classifier. We use RFM values because they are widely recognized as important indicators of customer value in e-commerce. *K*-means clustering is chosen for its ability to efficiently process large datasets (Agrawal et al., 2023). The XGBoost classifier is selected for its scalability and reliability in handling large, non-linear, high-dimensional datasets (Hajek et al., 2023).

Our approach is evaluated against industry standard practices, demonstrating superior performance in accuracy, precision, recall, and *F*1-score for loyalty prediction, as well as higher silhouette scores and lower Davies–Bouldin indices for customer segmentation. We also provide insights into the characteristics of different customer segments and their loyalty levels. The structure of the paper includes related research, the proposed hybrid approach, experimental setup, results and discussions, implications, conclusion, and suggestions for further research. This research is particularly pertinent given current trends in e-commerce, where understanding customer behavior and boosting loyalty is critical for maintaining market growth and competition.

Related Work

In this section, we review recent methods for customer segmentation and loyalty prediction in e-commerce, focusing on studies conducted between 2010 and 2024. This review aims to highlight the most significant research and identify existing gaps.

Customer Segmentation

One of the most crucial practices in marketing and e-commerce is customer segmentation, that is, the classification of customers into similar groups based on alike attributes or behaviors (Agrawal et al., 2023). There are segmentation techniques that include geographic, psychographic, behavioral, and demographic ones based on factors such as age, gender, income, location, lifestyle, personality quantity, frequency, and preferred products purchased.

Customer segmentation periods include recency (how recently a person purchased a product), frequency (how frequently a person purchases a product), and monetary value (which includes the magnitude of the purchase). Customers are evaluated in terms of purchase frequency, the total amount spent, and the period since the last purchase (Jauhar et al., 2024; Wan et al., 2022). Customers are divided into needy and unprofitable or maintained and loyal groups based on their RFM scores (Wan et al., 2022).

K-means clustering is a method worth noting for all aspiring marketers interested in customer segmentation. In this example of unsupervised learning models, data consists of clusters that are built based on the similarity between the data points and the center point of the cluster (Agrawal et al., 2023; Tabuena et al., 2022; Wan et al., 2022). *K*-means encompasses usage with other forms besides variables, such as RFM values and demographic variables. Aside from this, additional powerful techniques for consumer analysis include latent class analysis, self-organizing maps, and hierarchical clustering, to name a few. In as much as they meet their aim, all these approaches have disadvantages, such as high stability over time, high initialization sensitivity relations, and strong initialization relationship to outliers.

In response to previous remarks, some recent studies have discovered that ML approaches can be merged with classical models to improve segmentation performance. Lee and Jiang (2021) developed a method incorporating RFM and attitudinal factors that combine unsupervised clustering and supervised classification to ascertain customers' loyalty. Wu et al. (2022) advanced *K*-medoids enhancing by an adapted spare metric, penetration rate, improving the effectiveness and precision of customer segmentation.

Customer Loyalty Prediction

Predicting customer loyalty is important in appreciating loyal customers, enhancing verbal advertising, decreasing consumer distribution, or increasing purchase value per consumer over their lifetime (Heilman & Bowman, 2002). Recent studies have instead used different ML techniques to relate to loyalty. Prediction of customer loyalty is predominantly done using logistic regression. It predicts the presence of a binary dependent variable of “loyal” or “unloyal” with a dependent variable consisting of the independent variables (features) by applying a logistic curve (Myburg, 2023; Schapire, 2003). This technique can handle both categorical and numerical features. Decision trees are, without a doubt, the second-best option. They construct a represented figure in the shape of a tree where the nodes represent the features, the branches the features values, and the leaves are the class labels, which in this case can be loyal or not loyal (Huyut & Üstündağ, 2022; Kushwah et al., 2022). Decision trees can work with different feature types, as well as with missing data and outliers (Costa & Pedreira, 2023).

Other approaches include neural networks, support vector machines, random forests, and gradient boosting. Advanced clustering and segmentation techniques have been investigated to improve customer loyalty prediction. Othayoth and Muthalagu (2022) used an agglomerative clustering algorithm, while Ullah et al. (2023) analyzed a large e-commerce dataset using RFMT (Recency, Frequency, Monetary, and Time) with *k*-means, Gaussian, and DBSCAN. Tiwari et al. (2018) used self-organizing maps to classify purchasing behavior and supervised learning strategies such as nearest neighbor and help vector devices.

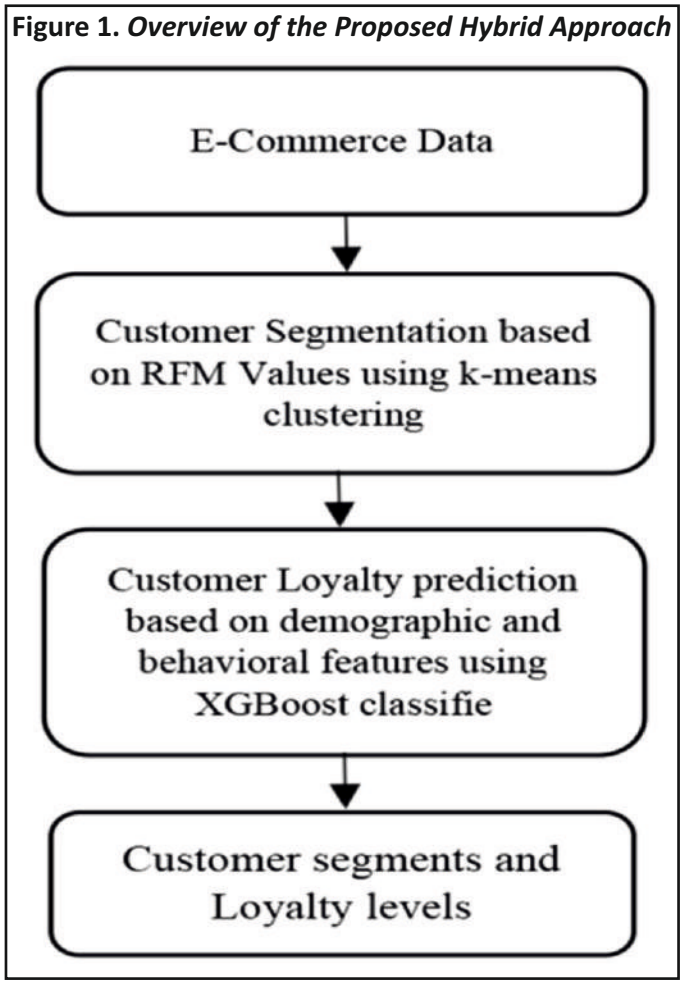
Research Gap

While numerous strategies for customer segmentation and loyalty prediction have been explored, there is a need for a strong, scalable hybrid technique that efficaciously integrates conventional and advanced ML techniques for complete customer segmentation and loyalty prediction. This paper pursuits to cope with this gap by presenting a hybrid technique combining statistical and ML techniques, in particular, the usage of RFM values with *k*-means clustering for segmentation and demographic and behavioral features with the XGBoost classifier for loyalty prediction.

Proposed Hybrid Approach

In this section, we go through the proposed hybrid methodology for e-commerce customer segmentation and loyalty prediction. Customer segmentation based on RFM criteria using *K*-means clustering and customer loyalty estimation based on demographic and behavioral characteristics using the XGBoost classifier are two main components of our framework. Figure 1 provides an overview of this approach, which shows the basic steps of customer segmentation.

Figure 1. ***Overview of the Proposed Hybrid Approach***



Customer Segmentation Based on RFM Values Using K-Means Clustering

The first part of our proposed framework involves clustering customers who will use *K* methods based on their RFM objectives. RFM values are important indicators of customer value in e-commerce. Recency refers to the time since the last purchase, frequency refers to the quantity purchased in a particular period, and monetary values specify the total amount spent during that period and the cash value (Myburg, 2023; Wan et al., 2022).

We calculate these values using the following assumptions.

- **Recency :** Current Date – Last purchased date.
- **Frequency :** The number of items ordered by the customer in the last 12 months.
- **Money Value :** The total amount of money the customer spends in the last 12 months.

These values are normalized using min–max scaling to standardize from 0 to 1. Customers are then divided into *K* groups using *K*-means clustering, an effective unsupervised learning algorithm (Rizki et al., 2020). The elbow method suitable for large-scale cases, which involves comparing the sum of squared error (SSE) for different values of *K* is used to determine the optimal number of clusters by identifying the point where the SSE curve bends sharply.

Customer Loyalty Prediction Based on Demographic and Behavioral Features Using XGBoost Classifier

The second component of our strategy utilizes the XGBoost classifier to predict customer loyalty based on demographic and behavioral characteristics (Hajek et al., 2023). Loyalty is defined as the likelihood of a customer making additional purchases within three months following their last purchase (Ajina, 2019). We use a binary indicator to determine whether or not a customer is loyal. Features studied include age, gender, location, product category, order frequency, order value, and order recency.

XGBoost, a powerful supervised learning algorithm, is used to estimate customer loyalty. XGBoost transforms several weak learners (decision trees) into a strong learner through an ensemble method that iteratively adds new trees to minimize a loss function (Chen et al., 2014). It is chosen for its scalability, accuracy, and ability to handle high-dimensional, non-linear data (Myburg, 2023). We assess the XGBoost classifier's performance using 10-fold cross-validation, employing accuracy, precision, recall, and *F*1-score as evaluation metrics. Comparisons are made with industry-standard techniques such as logistic regression, decision trees, neural networks, and support vector machines.

Research Methodology

This research adopts a quantitative approach, utilizing statistical and ML methods for data analysis. The study design involves the collection and analysis of a real-world e-commerce dataset from India containing customer orders from January 2019 to December 2022. The dataset included three files: list_of_orders.csv, which contains order ID, order date, and customer ID; order_details.csv, which includes order ID, product ID, and product category; and sales_target.csv which provides month-year combinations and sales target amounts. We employ a non-probability sampling technique, focusing on a sample frame of 9,514 customers who placed 50,617 orders across 10 product categories. This sampling method is chosen to capture a diverse and representative customer base within the Indian e-commerce market, ensuring the generalizability and applicability of our findings.

Data analysis is conducted using Python, leveraging libraries such as pandas for data manipulation, scikit-learn for clustering, and XGBoost for classification. We ensure the reliability of the scales used through

internal consistency checks, with a focus on maintaining high-reliability values, specifically aiming for a Cronbach's alpha greater than 0.7 to ensure robust and reliable measurements. The analysis spans the period from January 2019 to February 2022 and focuses on the Indian e-commerce industry, providing valuable insights into customer segmentation and loyalty estimates in this particular geographical and market context.

Analysis and Results

Customer Segmentation Based on RFM Values Using K-Means Clustering

At the beginning of our approach, we used *K*-means clustering to classify customers based on their RFM criteria. These parameters were calculated using the parameters shown in the section "Customer Segmentation Based on RFM Values Using *K*-Means Clustering" and normalized using min–max scaling. The optimal number of clusters was determined by the elbow method, which involves plotting the total squared error (SSE) against different values of *K* and identifying the point where the curve turns sharply, as shown in Figure 2. The SSE curve bends sharply at $K = 4$, indicating that the four clusters are optimal.

Figure 2. ***Elbow Plot for K-Means Clustering***

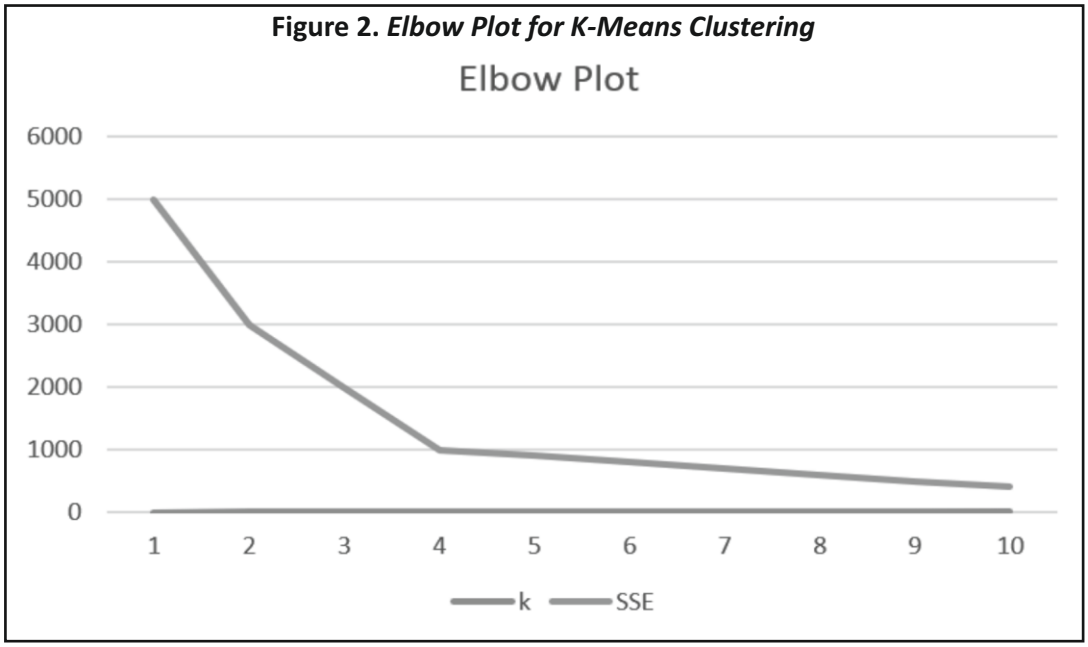


Table 1. ***Silhouette Score and Davies–Bouldin Index for K–Means Clustering with K = 4***

K	Silhouette Score	Davies–Bouldin Index
1	0.37	1.15
2	0.41	1.12
3	0.42	0.95
4	0.45	0.87
5	0.43	0.91
6	0.41	0.98

We then used *K*-means clustering with $K = 4$ for the segment customers. Clusters were assessed using the Silhouette score and the Davis–Boldin index. The Silhouette score, which ranges from –1 to 1, measures the similarity of the object to its category compared to other categories, with higher values indicating better clustering (Singh et al., 2022; Xiao et al., 2017); The Davis–Boldin index measures the separation between groups, with very low standard clustering means (Thomas et al., 2013).

Table 1 presents the silhouette score and Davies–Bouldin index for $K = 4$, showing that the clustering achieved a high silhouette score and a low Davies–Bouldin index, indicating good segmentation.

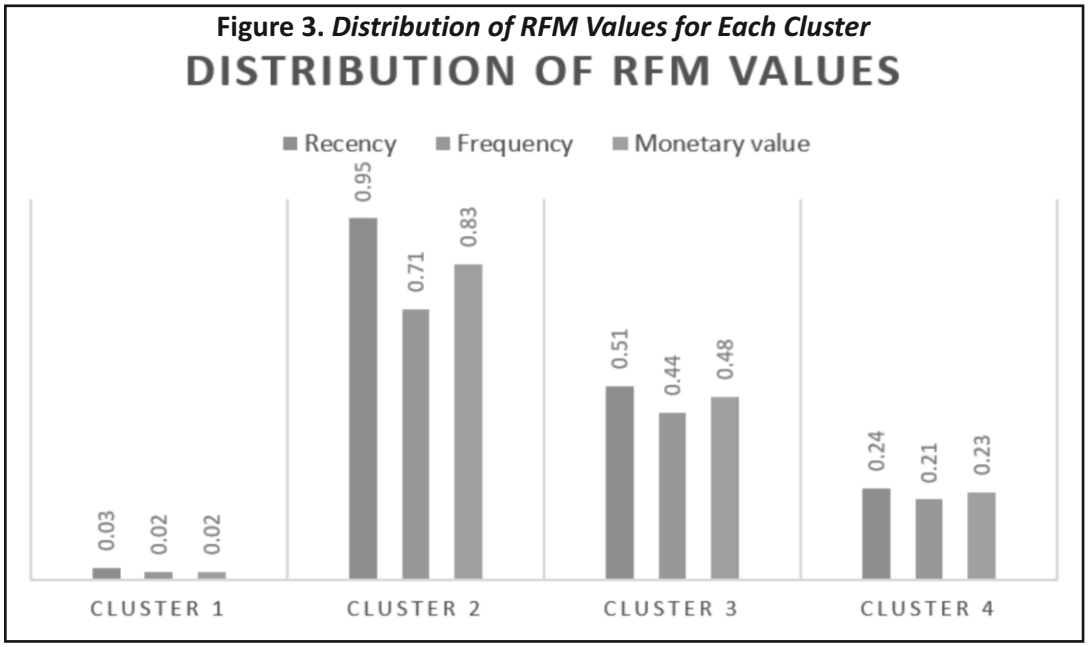
We further analyzed the characteristics of each cluster based on their RFM values, as shown in Table 2. Cluster 1 has the lowest recency, frequency, and monetary value; Cluster 2 has the highest values for these metrics; Cluster 3 has moderate values; and Cluster 4 has low recency but moderate frequency and monetary value. Figure 3 illustrates the distribution of RFM values for each cluster. Based on these observations, we labeled the clusters as follows:

- **Cluster 1 :** Lost customers
- **Cluster 2 :** Loyal customers
- **Cluster 3 :** Regular customers
- **Cluster 4 :** New customers

Table 2. ***Descriptive Statistics of Each Cluster***

Cluster	Recency	Frequency	Monetary Value	Number of Customers
Cluster 1	0.03	0.02	0.02	2,643
Cluster 2	0.95	0.71	0.83	1,586
Cluster 3	0.51	0.44	0.48	3,171
Cluster 4	0.24	0.21	0.23	2,114

Figure 3. ***Distribution of RFM Values for Each Cluster***



Customer Loyalty Prediction Based on Demographic and Behavioral Features Using XGBoost Classifier

The second stage of our approach involves predicting customer loyalty using the XGBoost classifier. Loyalty was defined as the likelihood of a customer making additional purchases within three months following their last purchase (Leninkumar, 2017), labeled as loyal or not loyal. We considered various demographic and behavioral features such as age, gender, location, product category, frequency of orders, order value, and recentness of orders.

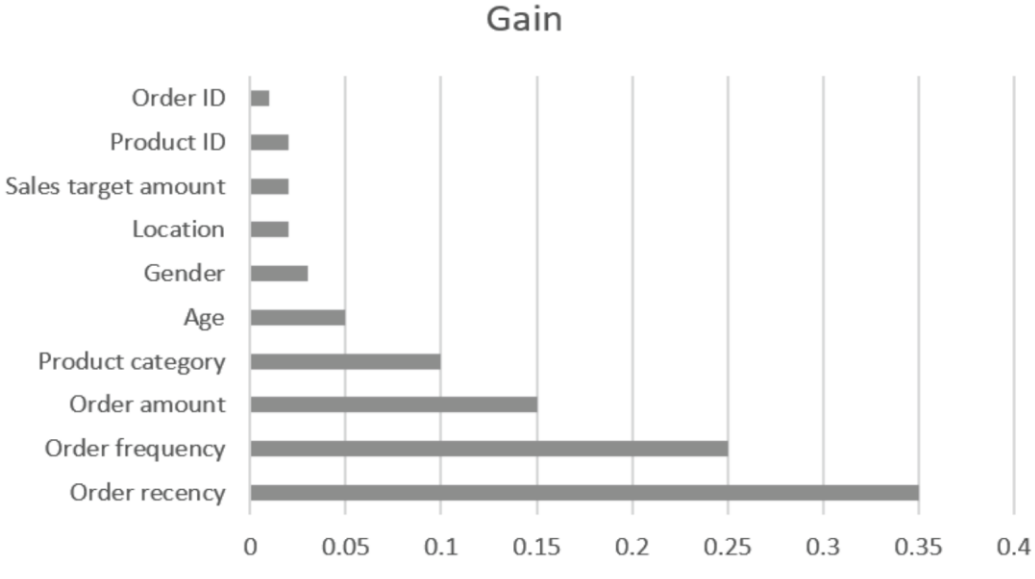
XGBoost, a scalable and reliable supervised learning algorithm, was chosen for its ability to handle high-dimensional data with non-linear relationships. We assessed the classifier's effectiveness using 10-fold cross-validation and evaluation metrics such as accuracy, precision, recall, and *F*1-score. Table 3 shows that the XGBoost classifier beat other industry-standard techniques, such as logistic regression, decision trees, neural networks, and support vector machines, in all evaluation measures.

Figure 4 shows how we used XGBoost to determine the relevance of features. Order recency was the most crucial factor, followed by order frequency, order amount, product category, age, gender, location, sales target amount, product ID, and order ID. This indicates that recent and frequent purchases, higher spending, preferences for specific product categories, and certain demographic characteristics significantly contribute to customer loyalty.

Table 3. ***Performance of Different Methods for Customer Loyalty Prediction***

Method	Accuracy	Precision	Recall	*F*1-score
Logistic regression	0.71	0.69	0.75	0.76
Decision tree	0.78	0.79	0.78	0.81
Neural network	0.81	0.83	0.79	0.82
Support vector machine	0.83	0.85	0.81	0.84
XGBoost classifier	**0.85**	**0.86**	**0.83**	**0.85**

Figure 4. ***Top 10 Features Ranked by Their Gain in XGBoost Classifier***



Implications

Managerial Implications

Our approach offers significant benefits for e-commerce businesses by providing deeper insights into customer behavior through segmentation based on RFM values and loyalty levels. This segmentation enables companies to tailor their pricing strategies, resource allocation, marketing campaigns, and product offerings to different market segments. Customized marketing strategies can be developed to target loyal customers with discounts and free shipping, engage regular customers with cross-selling opportunities, attract new customers with welcome emails and social proof, and re-engage lost customers with win-back offers. Predicting customer loyalty enables firms to improve retention and acquisition efforts, identify important clients, reduce churn, and increase recommendations, resulting in better-informed decision-making and optimized customer relationship management.

Theoretical Implications

Our hybrid approach, which combines *k*-means clustering and XGBoost classifier, contributes to the literature by demonstrating the effectiveness of combining statistical and ML techniques for customer segmentation and loyalty prediction. It outperforms other methods in accuracy, recall, precision, and *F*1-score for loyalty prediction and obtains a higher silhouette score and lower Davis–Bouldin index for customer segmentation, indicating better clustering quality. Furthermore, the XGBoost classifier's feature relevance ratings that have an impact on customer loyalty improve theoretical comprehension by identifying essential demographic and behavioral traits.

Policy Implications

Policymakers within e-commerce companies can utilize these insights for strategic resource allocation, directing sources greater efficiently to consumer segments primarily based on their cost and loyalty tiers. This involves growing focused marketing and customer engagement guidelines to focus efforts on segments with the highest potential for retention and growth. Establishing frameworks for non-stop performance assessment of customer segmentation and loyalty prediction models, incorporating clear metrics and feature importance insights can be vital. Furthermore, developing procedures for the ethical use and control of consumer information can protect privacy and compliance while also using data for improved business outcomes.

Conclusion

In this paper, we introduce a hybrid methodology for predicting customer loyalty and segmenting customers in e-commerce, where *K*-means clustering and XGBoost classifier are used. Integrating statistical and ML techniques, our method segments customers based on RFM criteria and loyalty through demographic and behavioral characteristics. We validated our approach by using a real-world Indian e-commerce dataset and benchmarking it against numerous industry standards. Our method shows good performance in terms of accuracy, precision, recall, and *F*1 scores for predicting loyalty and obtained high silhouette scores and low Davis–Bouldin indices for customer segmentation. We also provide insights into client segment quality and loyalty levels. This method provides several benefits to e-commerce businesses, including increased customer knowledge, the ability to tailor advertising campaigns and product offerings to specific segments, optimized resource allocation and pricing strategies, the identification of most valuable customers, multiplied word-of-mouth recommendations, lower churn costs, and improved decision-making and overall performance.

Limitations of the Study

Our hybrid approach faces numerous boundaries and challenges that need to be addressed in future research. First, our method relies on historical information to segment customers and to predict loyalty, but customer behavior and preferences may change over time because of various factors which include market trends, seasonality, and opposition. Therefore, our approach needs to be up to date, often with fresh records, to seize the dynamic nature of customer behavior and loyalty. Second, our technique uses a binary label (loyal or not loyal) to indicate customer loyalty primarily based on whether they made at least one buy in the three months following their latest buy.

This definition might not fully capture the nuances of customer satisfaction and retention. Customers who make a single buy in the following three months might not be as loyal as those who make numerous purchases during the same duration. Therefore, our approach calls for extra superior consumer loyalty metrics, inclusive of the net promoter score (NPS), consumer satisfaction score (CSAT), or customer effort score (CES). Third, our technique classifies customers primarily based on demographic and behavioral traits, but these functions will not fully constitute the variety of customer traits or behaviors. Features consisting of mental or emotional traits, together with persona trends, motivations, attitudes, and feelings, will not be pondered. Therefore, our approach should incorporate greater detailed features, which include sentiment analysis or psychographic functions.

Scope for Future Research

Future research could expand our hybrid technique by way of incorporating numerous key factors. Temporal characteristics, together with seasonality, fashion, and cycle, can be blanketed to seize the cyclical styles of customers. We will also include social capabilities such as network structure, influence dispersion, and word-of-mouth impacts to better understand consumer interactions and their impact on purchasing decisions and brand loyalty. Contextual functions such as product availability, pricing modifications, and promotional activities will be used to capture external impacts on consumer behavior and loyalty. Future studies can also use text-mining techniques, such as subject matter modeling, sentiment analysis, and emotion evaluation, to extract meaningful data from unstructured text facts like product descriptions and consumer reviews. Furthermore, deep knowledge of techniques, such as convolutional neural networks, can be used to assess consumer happiness and discontent with services or products. We anticipate that our hybrid technique may provide a powerful foundation for client segmentation and e-commerce loyalty prediction while also motivating more research in this field.

Authors' Contribution

Elamurugan Balasundaram conceptualized the study and led the project. P. Aranganathan designed the methodology and performed data analysis. R. Sivakumar collected the data and contributed to the interpretation of the results. Mathiazhagan Arumugam was responsible for the literature review and theoretical framework. A. Vinoth assisted in data collection and manuscript preparation. Krishna Sudhir Annavajjala contributed to the data analysis and provided inputs on manuscript revisions. All authors reviewed and approved the final manuscript.

Conflict of Interest

The authors declare that there is no conflict of interest regarding the publication of this article.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Acknowledgments

The authors would like to express their gratitude to their colleagues for their valuable feedback and support throughout this research project. Special thanks to the participants who contributed their time and insights, making this study possible.

References

Adelaar, T., Bouwman, H., & Steinfield, C. (2004). Enhancing customer value through click-and-mortar e-commerce: Implications for geographical market reach and customer type. *Telematics and Informatics*, *21*(2), 167–182. https://doi.org/10.1016/S0736-5853(03)00055-8

Agrawal, A., Kaur, P., & Singh, M. (2023). Customer segmentation model using K-means clustering on e-commerce. In *2023 International Conference on Sustainable Computing and Data Communication Systems (ICSCDS)* (pp. 1–6). IEEE. https://doi.org/10.1109/icscds56580.2023.10105070

Ahmed, N., & Kumari, A. (2022). The implication of e-commerce emerging markets in the post-Covid era. *International Journal of Entrepreneurship and Business Management*, *1*(1), 21–31. https://doi.org/10.54099/ijebm.v1i1.102

Ajina, A. S. (2019). The role of content marketing in enhancing customer loyalty: An empirical study on private hospitals in Saudi Arabia. *Innovative Marketing*, *15*(3), 71–84. https://doi.org/10.21511/im.15(3).2019.06

Andersson, S., & Börjeson, M. J. (2023). *Customer journey management within B2B e-commerce: A case study on how to implement customer journey management* (Report No. E2023:046). Chalmers Open Digital Repository. http://hdl.handle.net/20.500.12380/306160

Chen, T., He, T., Benesty, M., Khotilovich, V., Tang, Y., Cho, H., Chen, K., Mitchell, R., Cano, I., Zhou, T., Li, M., Xie, J., Lin, M., Geng, Y., Li, Y., & Yuan, J. (2014). Xgboost: Extreme gradient boosting [dataset]. In *CRAN: Contributed Packages*. The R Foundation. https://doi.org/10.32614/cran.package.xgboost

Costa, V. G., & Pedreira, C. E. (2023). Recent advances in decision trees: An updated survey. *Artificial Intelligence Review*, *56*(5), 4765–4800. https://doi.org/10.1007/s10462-022-10275-5

Dhote, T., & Zahoor, D. (2017). Framework for sustainability in e-commerce business models: A perspective based approach. *Indian Journal of Marketing*, *47*(4), 35–50. https://doi.org/10.17010/ijom/2017/v47/i4/112681

Hajek, P., Abedin, M. Z., & Sivarajah, U. (2023). Fraud detection in mobile payment systems using an XGBoost-based framework. *Information Systems Frontiers*, *25*, 1985–2003. https://doi.org/10.1007/s10796-022-10346-6

Heilman, C. M., & Bowman, D. (2002). Segmenting consumers using multiple-category purchase data. *International Journal of Research in Marketing*, *19*(3), 225–252. https://doi.org/10.1016/s0167-8116(02)00077-0

Huyut, M. T., & Üstündağ, H. (2022). Prediction of diagnosis and prognosis of COVID-19 disease by blood gas parameters using decision trees machine learning model: A retrospective observational study. *Medical Gas Research*, *12*(2), 60–66. https://doi.org/10.4103/2045-9912.326002

Jauhar, S. K., Chakma, B. R., Kamble, S. S., & Belhadi, A. (2024). Digital transformation technologies to analyze product returns in the e-commerce industry. *Journal of Enterprise Information Management, 37*(2), 456–487. https://doi.org/10.1108/jeim-09-2022-0315

Joshi, D., & Achuthan, S. (2016). A study of trends in B2C online buying in India. *Indian Journal of Marketing*, *46*(2), 22–35. https://doi.org/10.17010/ijom/2016/v46/i2/87248

Kumar, A., Gupta, S. L., & Kishor, N. (2016). The antecedents of customer loyalty: Attitudinal and behavioral perspectives based on Oliver's loyalty model. *Indian Journal of Marketing*, *46*(3), 31–53. https://doi.org/10.17010/ijom/2016/v46/i3/88996

Kushwah, J. S., Kumar, A., Patel, S., Soni, R., Gawande, A., & Gupta, S. (2022). Comparative study of regressor and classifier with decision tree using modern tools. *Materials Today: Proceedings*, *56*(Part 6), 3571–3576. https://doi.org/10.1016/j.matpr.2021.11.635

Lee, H. F., & Jiang, M. (2021). A hybrid machine learning approach for customer loyalty prediction. In H. Zhang, Z. Yang, Z. Zhang, Z. Wu, & T. Hao (eds.), *Neural computing for advanced applications. NCAA 2021. Communications in computer and information science* (Vol. 1449, pp. 211–226). Springer. https://doi.org/10.1007/978-981-16-5188-5_16

Leninkumar, V. (2017). The relationship between customer satisfaction and customer trust on customer loyalty. *International Journal of Academic Research in Business and Social Sciences*, *7*(4), 450–465. https://doi.org/10.6007/IJARBSS/v7-i4/2821

Myburg, M. E. (2023). *Using recency, frequency and monetary variables to predict customer lifetime value with XGBoost*. Faculty of Science, Department of Computer Science. http://hdl.handle.net/11427/38088

Nisar, T. M., & Prabhakar, G. (2017). What factors determine e-satisfaction and consumer spending in e-commerce retailing? *Journal of Retailing and Consumer Services, 39*, 135–144. https://doi.org/10.1016/j.jretconser.2017.07.010

Othayoth, S. P., & Muthalagu, R. (2022). Customer segmentation using various machine learning techniques. *International Journal of Business Intelligence and Data Mining*, *20*(4), 480–496. https://doi.org/10.1504/IJBIDM.2022.123218

Rahayu, S., Cakranegara, P. A., Simanjorang, T. M., Syobah, S. N., & Arifin. (2022). Implementation of customer relationship management system to maintain service quality for customer. *Enrichment: Journal of Management*, *12*(5), 3856–3866. https://doi.org/10.35335/enrichment.v12i5.939

Rizki, B., Ginasta, N. G., Tamrin, M. A., & Rahman, A. (2020). Customer loyalty segmentation on point of sale system using recency-frequency-monetary (RFM) and K-means. *Jurnal Online Informatika*, *5*(2), 130–136. https://doi.org/10.15575/join.v5i2.511

Schapire, R. E. (2003). The boosting approach to machine learning: An overview. In D. D. Denison, C. C. Holmes, M. H. Hansen, B. Mallick, & B. Yu (eds.), *Nonlinear estimation and classification. Lecture notes in statistics* (Vol. 171, pp. 149–171). Springer. https://doi.org/10.1007/978-0-387-21579-2_9

Shobana, J., Gangadhar, C., Arora, R. K., Renjith, P. N., Bamini, J., & Chincholkar, Y. D. (2023). E-commerce customer churn prevention using machine learning-based business intelligence strategy. *Measurement: Sensors*, *27*, Article ID 100728. https://doi.org/10.1016/j.measen.2023.100728

Singh, A., Inamdar, A. G., Kaimal, A. R., Mahajan, V., & Priya, R. (2022). Chapter 5: Customization of product/service on e-commerce websites. In, *Changing face of e-commerce in Asia* (pp. 79–96). World Scientific Publishing. https://doi.org/10.1142/9789811245992_0005

Tabuena, A. C., Necio, S. M., Macaspac, K. K., Bernardo, M. P., Domingo, D. I., & De Leon, P. D. (2022). A literature review on digital marketing strategies and its impact on online business sellers during the COVID-19 crisis. *Asian Journal of Management, Entrepreneurship and Social Science*, *2*(01), 141–153. https://ajmesc.com/index.php/ajmesc/article/view/43

Thomas, J. C., Peñas, M. S., & Mora, M. (2013). New version of Davies-Bouldin Index for clustering validation based on cylindrical distance. In *2013 32nd International Conference of the Chilean Computer Science Society (SCCC)*. IEEE. https://doi.org/10.1109/sccc.2013.29

Tiwari, R., Saxena, M. K., Mehendiratta, P., Vatsa, K., Srivastava, S., & Gera, R. (2018). Market segmentation using supervised and unsupervised learning techniques for e-commerce applications. *Journal of Intelligent & Fuzzy Systems*, *35*(5), 5353–5363. https://doi.org/10.3233/jifs-169818

Ullah, A., Mohmand, M. I., Hussain, H., Johar, S., Khan, I., Ahmad, S., Mahmoud, H. A., & Huda, S. (2023). Customer analysis using machine learning-based classification algorithms for effective segmentation using recency, frequency, monetary, and time. *Sensors*, *23*(6), 3180. https://doi.org/10.3390/s23063180

Wan, S., Chen, J., Qi, Z., Gan, W., & Tang, L. (2022). Fast RFM model for customer segmentation. In *WWW '22: Companion proceedings of the web conference 2022* (pp. 965–972). ACM Digital Library. https://doi.org/10.1145/3487553.3524707

Wu, Z., Jin, L., Zhao, J., Jing, L., & Chen, L. (2022). Research on segmenting e-commerce customer through an improved K-Medoids clustering algorithm. *Computational Intelligence and Neuroscience*, *2022*. Article ID 9930613. https://doi.org/10.1155/2022/9930613

Xiao, J., Lu, J., & Li, X. (2017). Davies Bouldin Index based hierarchical initialization K-means. *Intelligent Data Analysis*, *21*(6), 1327–1338. https://doi.org/10.3233/ida-163129

About the Authors

Dr. Elamurugan Balasundaram has 17 years of teaching experience at several educational institutions. He is an Associate Professor at Sri Manakula Vinayagar Engineering College in Puducherry, India. Additionally, he received a diploma in a couple of managerial disciplines. His areas of interest in his research and teaching are the theory and implementation of marketing management, entrepreneurship, economics, and human resource practices. In addition to three book chapters, he has over nine articles published in international conferences and peer-reviewed journals.

Dr. P. Aranganathan is currently working as an Associate Professor at the Gnanam School of Business, Thanjavur, Tamil Nadu, India. He earned a Bachelor of Engineering (B.E.) in Mechanical and Production Engineering, as well as an MBA in Human Resource Management. He completed his Ph.D. in Management (HR) from Bharathiar University, India. He has authored four books (*Principles of Management, Operations Management, Logistics and Supply Management,* and *Total Quality Management*) and holds more than 30 publications in various international and national peer-evaluated indexed journals, proceedings, and books and two published patents. His areas of interest include strategic HRM and sustainable business practices.

Dr. Krishna Sudhir Annavajjala is a Professor and HOD of the Management Studies, Department of KLEF Hyderabad. He has made major contributions to academic business research based on his graduate teaching experience in management. He focuses on organizational behavior, strategic management, and business innovation. He has authored scores of research papers in leading national and international journals. He has been awarded for his dedication to teaching and coaching students in their academic and career pursuits.

Dr. R. Sivakumar is an Assistant Professor in the Department of Management Studies at Sri Manakula Vinayagar Engineering College (SMVEC), Puducherry, an autonomous institution. He received his Ph.D. from Anna University and his MBA from SMVEC Affiliated with Pondicherry Central University. Previously, he served as an Assistant Professor at Surya Group of Institutions in Villupuram. He has published six articles in Scopus and other indexed journals and has presented more than 20 research papers at various conferences. His research interests include customer experience management and digital marketing.

Mathiazhagan Arumugam is an Assistant Professor in the Department of Management Studies at Sri Manakula Vinayagar Engineering College (SMVEC) in Puducherry. He turned in his Ph.D. He completed his Ph.D. thesis at Annamalai University and his MBA at Mailam Engineering College (Anna University). Before that, he was a Lecturer at Loyola Institute of Technology in Chennai. His work has found recognition in Scopus and other indexed journals, with a total of two publications published while also presenting over seven research papers at different conferences. His research has focused on psychological capital and social support.

A. Vinoth, Ph.D., is an Assistant Professor/Department of Management Studies at Sri Manakula Vinayagar Engineering College (SMVEC) in Puducherry, which is an autonomous institution from January 2021. He received his Ph.D. in Green Marketing in June 2017. In addition, before joining SMVEC, he worked as an Assistant Professor at Vels University, Chennai. He has published in Scopus-indexed and other high-quality publications, as well as presented more than seven research papers at various conferences. He specializes in business economics, operations, and marketing.